A Meta-Analysis Approach for Feature Selection in Network Traffic Research

Daniel C. Ferreira, Félix Iglesias Vázquez, Gernot Vormayr,
Maximilian Bachl, Tanja Zseby

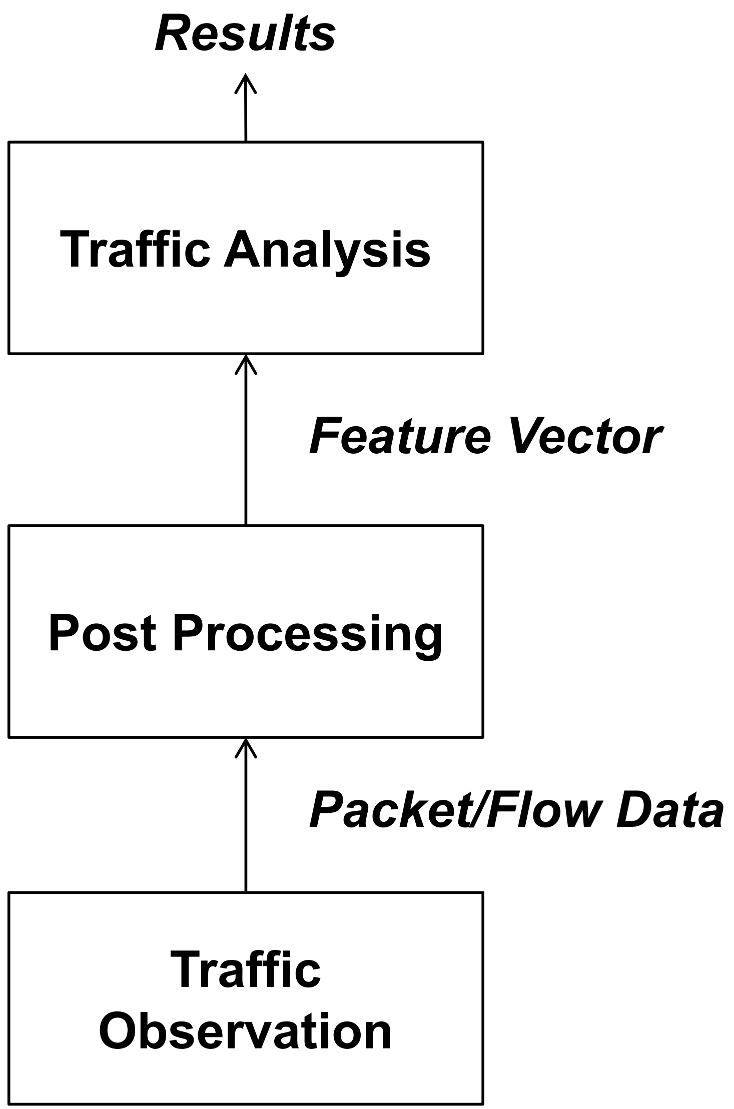
Institute of Telecommunications
TU Wien

Network Traffic Analysis

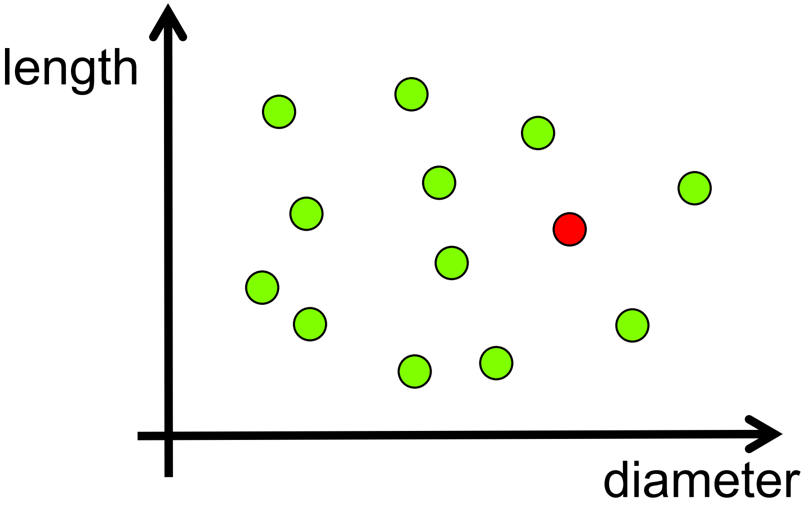
Feature Selection:
Select most suitable features

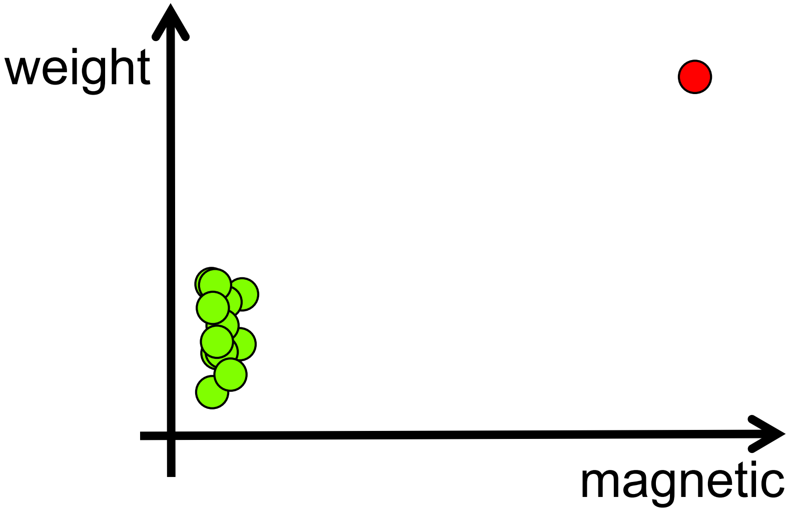
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \end{bmatrix}$$

🟩 🟥 🟩 🟩

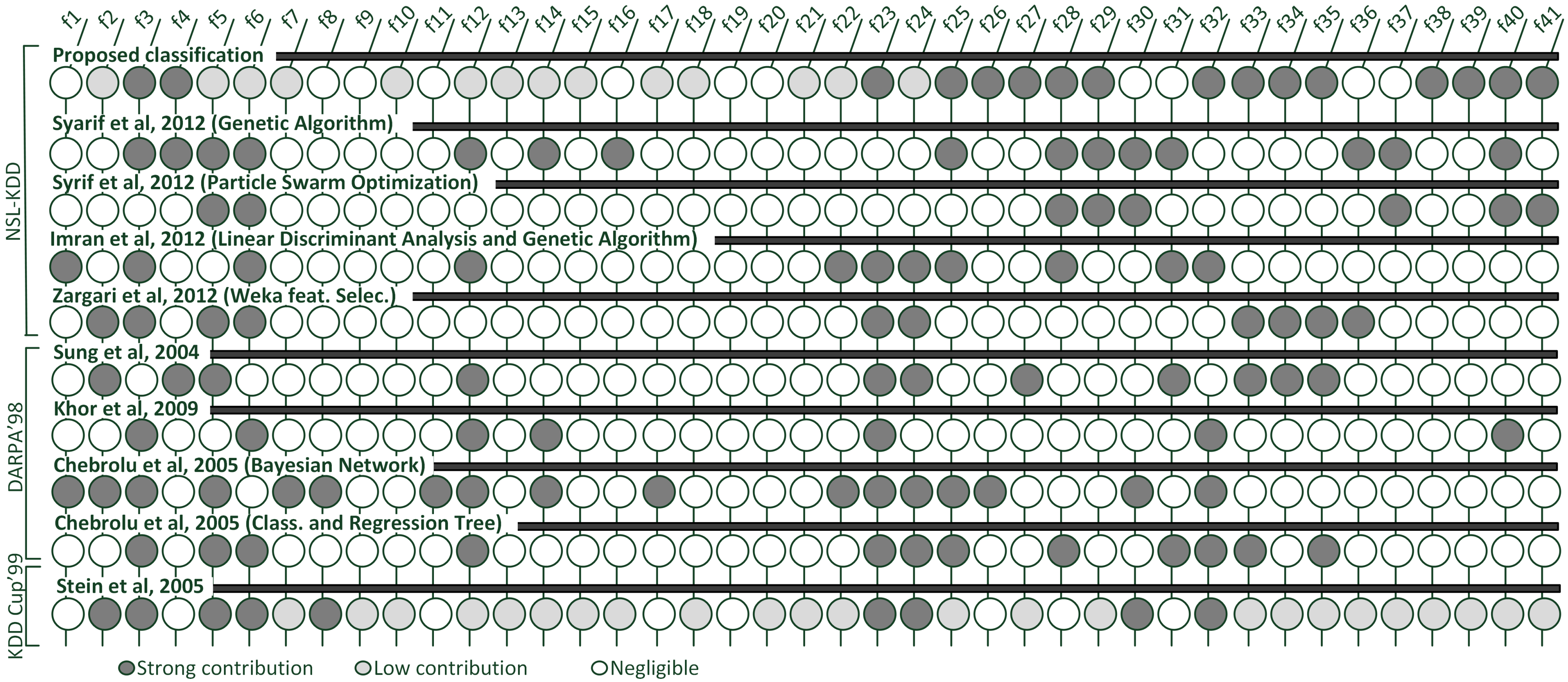


Well-chosen Features ➔ Simplified Analysis





Agree to Disagree



		f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17	f18	f19	f20	f21	f22	f23	f24	f25	f26	f27	f28	f29	f30	f31	f32	f33	f34	f35	f36	f37	f38	f39	f40	f41
NSL-KDD	Proposed classification	○	◐	●	●	◐	◐	◐	○	○	◐	○	◐	◐	◐	◐	○	◐	◐	○	○	◐	◐	●	◐	●	●	●	●	●	○	○	●	●	●	●	○	○	●	●	●	●
	Syarif et al, 2012 (Genetic Algorithm)	○	○	●	●	●	●	○	○	○	○	○	●	○	●	○	●	○	○	○	○	○	○	○	○	●	○	○	●	●	●	●	○	○	○	○	●	●	○	○	●	○
	Syrif et al, 2012 (Particle Swarm Optimization)	○	○	○	○	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●	●	○	○	○	○	○	○	●	○	○	●	●
	Imran et al, 2012 (Linear Discriminant Analysis and Genetic Algorithm)	●	○	●	○	○	●	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	●	●	●	●	○	○	●	○	○	●	●	○	○	○	○	○	○	○	○	○
	Zargari et al, 2012 (Weka feat. Selec.)	○	●	●	○	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●	○	○	○	○	○	○	○	○	●	●	●	●	○	○	○	○	○
DARPA'98	Sung et al, 2004	○	●	○	●	●	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	●	●	○	○	●	○	○	○	●	○	●	●	●	○	○	○	○	○	○
	Khor et al, 2009	○	○	●	○	○	●	○	○	○	○	○	●	○	●	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	●	○
	Chebrolu et al, 2005 (Bayesian Network)	●	●	●	○	●	○	●	●	○	○	●	●	○	●	○	○	●	○	○	○	○	●	●	●	●	●	○	○	○	●	○	●	○	○	○	○	○	○	○	○	○
	Chebrolu et al, 2005 (Class. and Regression Tree)	○	○	●	○	●	●	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	●	●	●	○	○	●	○	○	●	●	●	○	●	○	○	○	○	○	○
KDD Cup'99	Stein et al, 2005	○	●	●	○	●	●	◐	●	◐	◐	○	◐	◐	◐	◐	◐	○	◐	○	◐	◐	◐	●	●	◐	○	◐	○	◐	●	○	●	◐	◐	◐	◐	◐	◐	◐	◐	◐

● Strong contribution ◐ Low contribution ○ Negligible

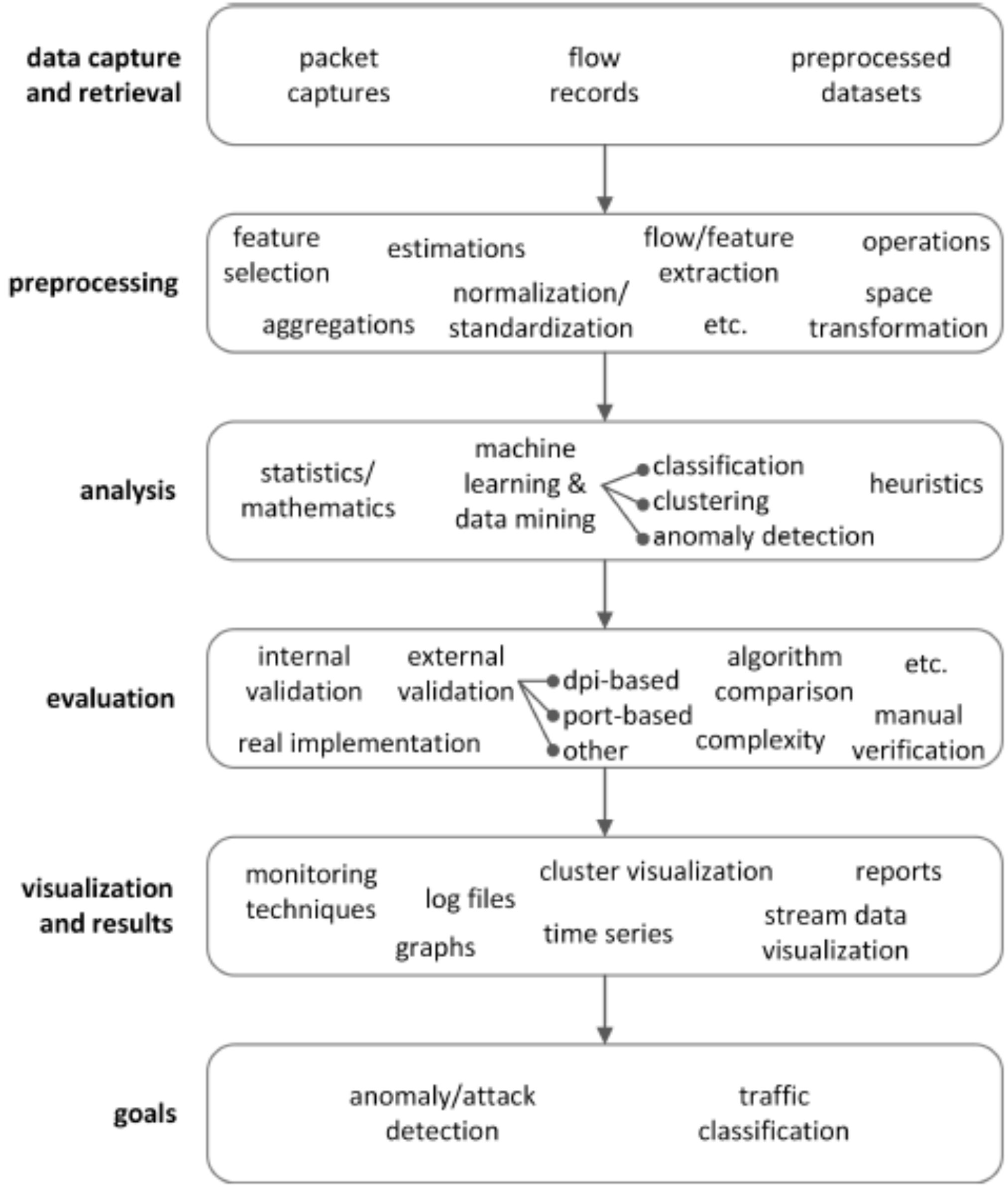
Source: Iglesias, Zseby: "*Analysis of network traffic features for anomaly detection*"; Machine Learning, **101** (2015), 1; 59 - 84.

Why a Meta Analysis?

- Meta-Analysis common in other disciplines
  - Structures the state of art
  - Combines existing results
  - Identifies agreements/disagreements in the community
  - Provides basis for gap analysis
- Provides information about
  - Availability of data and tools
  - Parameter settings
  - Validation Methods
  - Terminology and notation

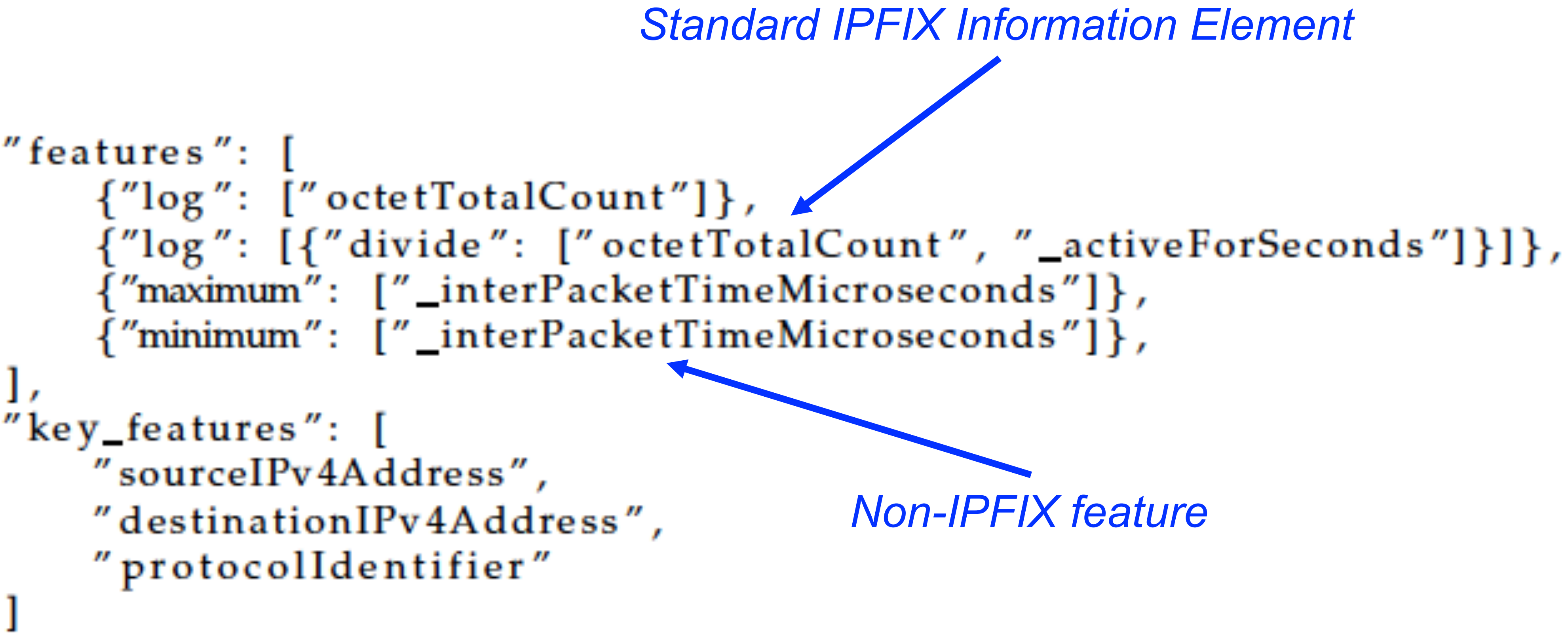
➔ Supports reproducibility and comparability

Data Structure



Example: Features

- Base features
- Operations on base features
- Flow keys

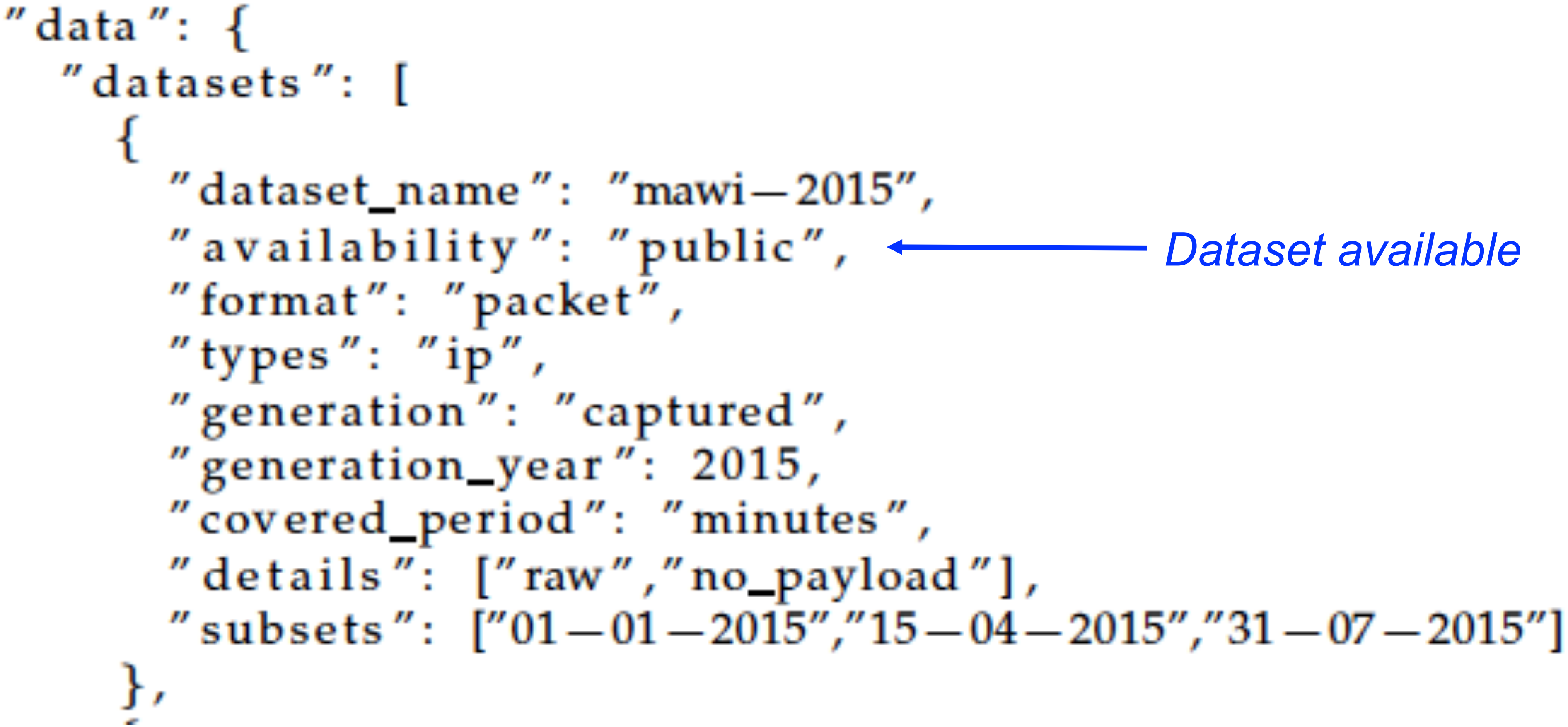


Standard IPFIX Information Element

```
"features": [
    {"log": ["octetTotalCount"]},
    {"log": [{"divide": ["octetTotalCount", "_activeForSeconds"]}]},
    {"maximum": ["_interPacketTimeMicroseconds"]},
    {"minimum": ["_interPacketTimeMicroseconds"]},
],
"key_features": [
    "sourceIPv4Address",
    "destinationIPv4Address",
    "protocolIdentifier"
]
```

Non-IPFIX feature

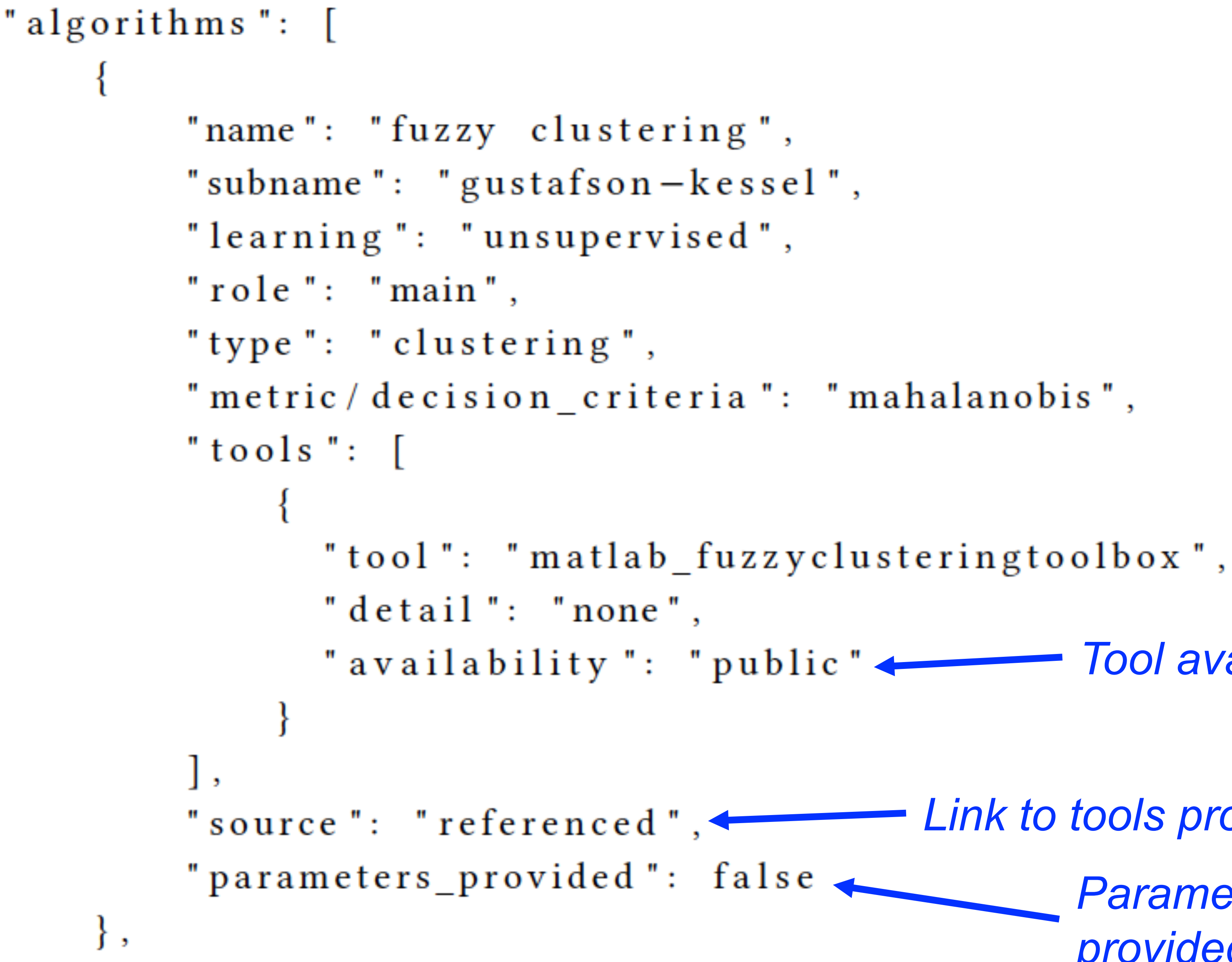
Example: Data Set



```
"data": {
  "datasets": [
    {
      "dataset_name": "mawi−2015",
      "availability": "public",
      "format": "packet",
      "types": "ip",
      "generation": "captured",
      "generation_year": 2015,
      "covered_period": "minutes",
      "details": ["raw","no_payload"],
      "subsets": ["01−01−2015","15−04−2015","31−07−2015"]
    },
```

Dataset available

Example: Algorithms



```
"algorithms": [
    {
        "name": "fuzzy clustering",
        "subname": "gustafson-kessel",
        "learning": "unsupervised",
        "role": "main",
        "type": "clustering",
        "metric/decision_criteria": "mahalanobis",
        "tools": [
            {
               "tool": "matlab_fuzzyclusteringtoolbox",
               "detail": "none",
               "availability": "public"
            }
        ],
        "source": "referenced",
        "parameters_provided": false
    },
```

Tool available

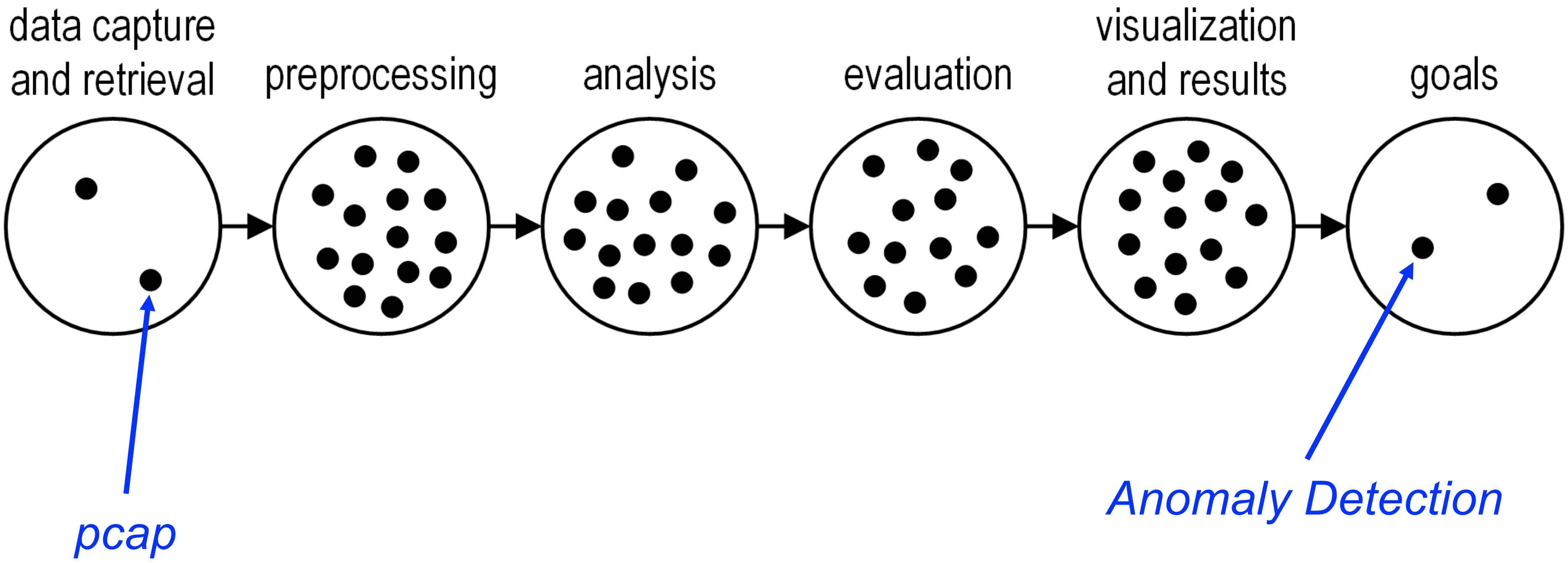
Link to tools provided

Parameters not provided

Initial Results

- 71 Papers from years 2005 to 2017

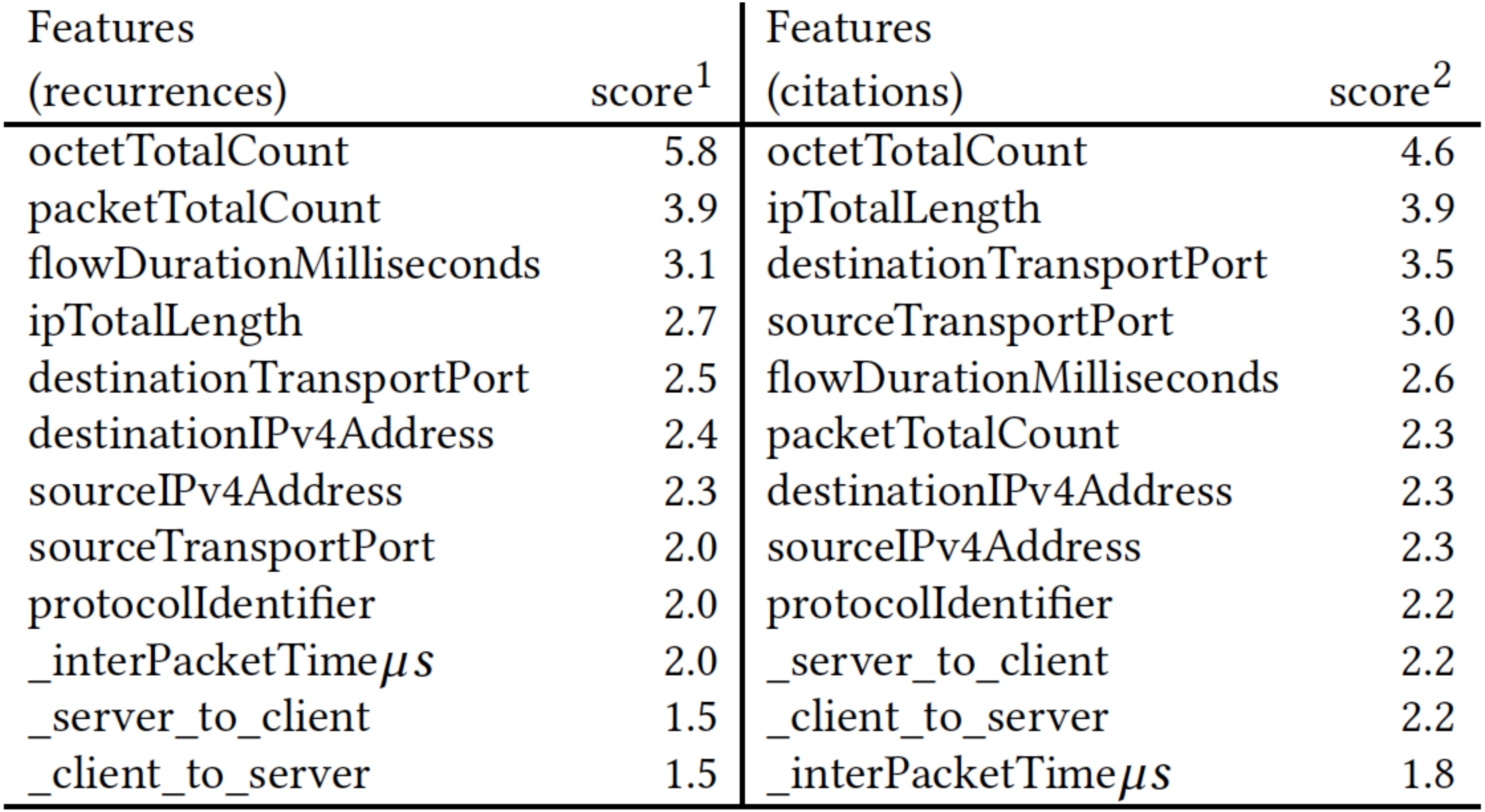
Analysis Chain



Initial Results

- Flow Definitions
  - 64.6% of papers that define a flow-key use classical 5-tuple {sIP, dIP, sPort, dPort, Protocol}
  - 70.8% use bi-directional flows
  - 83.1% use flow-based features
- Data Sets
  - 46.5% use at least one public data set
- Most Common Features
  - Number of papers that use a specific base feature
  - Number of papers weighted with their citations $\log_{10}$(citations)

Most Recurrent Base Features



Features		Features	
(recurrences)	score[1]	(citations)	score[2]
octetTotalCount	5.8	octetTotalCount	4.6
packetTotalCount	3.9	ipTotalLength	3.9
flowDurationMilliseconds	3.1	destinationTransportPort	3.5
ipTotalLength	2.7	sourceTransportPort	3.0
destinationTransportPort	2.5	flowDurationMilliseconds	2.6
destinationIPv4Address	2.4	packetTotalCount	2.3
sourceIPv4Address	2.3	destinationIPv4Address	2.3
sourceTransportPort	2.0	sourceIPv4Address	2.3
protocolIdentifier	2.0	protocolIdentifier	2.2
_interPacketTimeμs	2.0	_server_to_client	2.2
_server_to_client	1.5	_client_to_server	2.2
_client_to_server	1.5	_interPacketTimeμs	1.8

Summary

- Meta Analysis for Network Traffic Analysis
  - Supports comparability and reproducibility
  - Focus on feature selection, but much more information collected
- JSON files
  - Structured, searchable state of art
  - Fast extraction of relevant information from papers
- Initial results
  - Most common features
  - Flow definitions
  - Usage of public data sets
- → Data allows for many further analysis opportunities

Discussion

- Manual data curation ➔ Errors
  - Involve authors (check and correct)
- Analysis just shows “preferred” features, methods
  - ➔ not necessarily the best!
- Incentives to fill data base
  - Conferences can require to add accepted papers
  - Students can add data when exploring state of art
  - Searchable data base may increase citations for papers included

- All data, documentation, paper data base available at: ***www.cn.tuwien.ac.at/meta***

Thank you!

Contact: tanja.zseby@tuwien.ac.at